This document has been approved
for public release and sale;
its distribution is unlimited.

AD 666 228

TECHNICAL REPORT
68-10-PR

STUDIES ON ACCEPTANCE TESTING METHODOLOGY:
Preliminary Studies on Characteristics of Taste Panel, Sample Size, and Contrast and Convergence Effects

by

Joseph M. Kamen David R. Peryam
David B. Peryam Beverley J. Kroll

Peryam & Kroll Research Corporation
Chicago, Illinois 60645

Contract No. DA19-129-AMC-734 (N)

Project reference:
1T024701A121-02

July 1967

Pioneering Research Laboratory
U. S. ARMY NATICK LABORATORIES
Natick, Massachusetts 01760

FORWORD

In feeding the Armed Forces of the United States, increasing emphasis is directed toward providing foods of maximum acceptability within the constraints of specialized military requirements. Within this context, acceptable foods are viewed as items which will be consumed, and consumed with a degree of pleasure.

Such emphasis has resulted in a greater need for research on sensory evaluation techniques. Accurate and economic psychophysical methods will provide the food scientist with the proper tools for predicting serviceman's acceptance of various foods.

The work covered in this report was performed by the Peryam & Kroll Research Corporation, under Contract No. DA19-129-AMC-734 (N) and represents only a segment of the original undertaking. Research on the topics of optimum number of samples to be served in a test and contrast and convergence effects comprise the major portion of this report. Official investigators were Joseph M. Kamen, David R. Peryam, David B. Peryam, and Beverley J. Kroll. All taste tests were performed at the U. S. Army Natick Laboratories under the direction of the Project Officer, Joel L. Sidel, Acceptance Group, Behavioral Sciences Division, Pioneering Research Laboratory.

G. P. Mandels
Acting Director
Pioneering Research Laboratory

APPROVED:

DALE H. SIELING, Ph.D.
Scientific Director

WILLIAM M. MANTZ
Brigadier General, USA
Commanding

CONTENTS

LIST OF TABLES

LIST OF TABLES (Continued) Page

ABSTRACT

This report summarizes research accomplished in methodology aspects of sensory evaluation testing. It also discusses certain studies which were designed, but, for administrative reasons, could not be completed.

Two main studies are presented in detail. The first investigated the effect of the number of samples upon differences in preference between selected samples as a function of whether they were included in the first half or second half of a series. There was no evidence that the number of samples — varying from 2 to 12 — had any consistent or significant effect on preference differences; however, the data suggest several hypothesis for future investigation.

The second study attempted primarily to determine the effect of a fresh *vs* an irradiated control on preference differences among various irradiated samples. There were logical inconsistencies in the data; however, there was no basis for concluding that a fresh control attenuates the differences in preference. It appeared that quality control of the test products needed to be tightened. Methological aspects of this study and related studies are analyzed.

The implications of these studies for sensory evaluation procedures and methological research in this area are discussed. Recommendations for further in-house work by the U.S. Army Natick laboratories are made.

The report also covers the topic of sampling test subjects in-house, describes the panel population in the U.S. Army Natick Laboratories, and points out certain interrelationships among the panel member's background characteristics. These data suggest certain points for future investigation.

Studies on Acceptance Testing Methodology: Preliminary Studies on Characteristics of Taste Panel, Samole Size, and Contrast and Convergence Effects

Chapter I. INTRODUCTION

This report is concerned with work under the contract from the proposal stage to its termination. The initial scope and proposal are first discussed, then research, both completed and not completed, is described. Finally, recommendations are made for further work in this area.

Scope

The initiai scope of this contract was divided into three topics:

1. Sampling. Investigate the differences in preference evaluation related to taste panel composition, using the volunteer population of military and civilian employees at the Natick Laboratories, which is somewhat diverse in regard to a variety of personal and demographic factors.

2. Methodology. Investigate the optimum number of samples to be included in a test, the effects of sample order, procedures of sample presentation, and the appropriateness of various measurement scales for different types of evaluation.

3. Sensory vs Instrumental Measurement. Study the relationship between sensory and instrumental measurements of taste and odor, where the instrumental measurements might include chemical, textural, temperature, or color classifications.

At the outset, it was recognized that the scope was extremely broad and that much more work could be planned than could be conducted, analyzed, and reported. In particular, it seemed unlikely that time or funds would permit any extensive work on the third topic.

Approaches

Early meetings between the Contractor's personnel and the Project Officer were devoted to outlining and specifying in concrete terms the types and priority of research, and the delineation of areas of responsibility. The main emphasis was placed upon the first two topics.

The background of the work on the first topic (Sampling), and the results accomplished prior to its termination are covered in Chapter II.

A series of five experiments was proposed in connection with the second topic (Methodology) and detailed designs and analyses were developed.

1. Determination of limits of number of samples. The purpose was to determine the effect of extended testing on preference differentiation among samples. If longer test sessions yield as reliable results as do short ones, experimental designs can be employed which provide for a greater number of samples per subject, thereby increasing efficiency.

2. Procedural methods for increasing efficiency. The purpose was to ascertain whether the number of samples which can be effectively evaluated can be increased by certain procedural changes, in particular, whether a "rest-break" between segments of four samples each can overcome loss of preference discrimination — if such loss occurs. Such procedural changes would result in testing economies.

3. Balanced incomplete block (B.I.B) vs single classification designs.

   The purpose was to compare these designs, primarily on the basis of relative efficiency and, secondarily, for similarity of results (e.g., consistency of rank order of preference and spread of mean ratings) according to the number of samples evaluated. If the interpretations derived from B.I.B.'s are essentially the same as those from complete single classification designs, within specified limits on the number of samples, then the frequency of use of B.I.B's could be reduced with savings in time and money.

4. Effect of testing different food types in a single test session.

The purpose was to find out whether the permissible number of samples per session could be increased by serving different types of foods at the same session. In effect, this would mean running two different tests concurrently. It again sought to determine whether the number of samples per subject could be increased without changing the overall conclusions.

5. Further study of contrast and convergence effects.

There was particular interest here in the interactive efrects which arise in testing series which contain both irradiated samples and fresh controls. The objective of such tests is to achieve efficient discrimination not only among the irradiated samples but also between the fresh control and the irradiated samples as a group. The problem was to determine what interactions occur and whether the two purposes are compatible.

Consultations with the Project Officer resulted in the following order of priority for these experiments: #1, #5, #4, #3, and #2. Only the first two were completed. The results of Experiment #1 are presented in Chapter III, and Experiment #5 is the subject of Chapter IV.

Delination of Responsibility

Sampling. The Natick laboratories administered the questionnaires to get background information on the Natick panels and obtained the frequency distributions of each background characteristic. The Contractor, in consultation with the Project Officer, selected the background characteristics, developed the questionnaire, and recommended the food types to be used in the follow-through investigation. It had been planned for the Natick Laboratories to prepare a master background card for each respondent, collate the taste test ratings with the master card as required, and perform routine analyses of data. The Contractor would have performed the analyses beyond the basic prescribed routines. Work on this topic was terminated soon after analysis of the background questionnaire.

Methodology. The Contractor provided the specific test designs and consulted with the Project Officer in regard to the test foods to be employed. The Project Officer conducted and supervised the taste-tests in the Natick Food Acceptance Laboratory and was responsible for the performance of preliminary analyses of the data by the Natick Computer Branch. The Contractor monitored the experiments, performed analyses beyond the basic routines, interpreted the results, and made recommendations.

Chapter II. SAMPLING

The purpose of this phase was to estimate the effects of population variables upon the preference evaluations of different types of foods. These estimates could be used (a) to determine the generalizability of taste test results to the ultimate user (military) population, and (b) to recommend changes in preference assessment methods and procedures to enhance generalizability and validity. Negative results — lack of significant relationships between preference and demographic factors — would indicate that food rating behavior is largely independent of those characteristics on which the ultimate user population might differ from the Natick panel. This would imply satisfactory generalizability with current operating methods and procedures.

A secondary purpose was to determine whether, for various food types, a "self-selection" factor exists in taste-test sessions. Panel members sometimes refuse to make themselves available to test certain kinds of foods; and others may be eager to evaluate foods which they consider particularly appealing. This factor might bias the results. The first problem was to determine whether and, if so, under what conditions "self-selection" arises. The second was to ascertain whether "self-selection" does affect the interpretation of taste-test results.

1. Plan of Approach

The plan was to administer a background information questionnaire to all members of the Natick panel, analyze the frequency distributions, then select a limited number of background characteristics for study on how they were related to preferences within certain product categories.

The steps were as follows:

(a) Administer questionnaires to obtain information about: age, sex, state in which the participant spent most of his first 16 years, education, military service, years in Federal service (including military), subsistence on Army garrison rations continuously for one month or longer, subsistence on Army operational rations for three days or longer, attitude toward participating in the taste-test program, and opinion about the value of the taste-test activity. In addition, the number of taste-tests in which each person had particioated was obtained from the laboratory records.

(b) Obtain frequency distributions on each questionnaire item.

(c) Determine which factors were most feasible and worthwhile to investigate, establish two categories for each factor to obtain the best 50-50 split, and recode the information.

(d) Revise each participant's identification card to include the coded background information. Assign a new identification number to each subject and collate the background information with the ratings for each sample in tests meeting certain criteria.

(e) Select certain product categories for investigation. The following had been tentatively selected: irradiated meats, space foods, margarine and cheese, soups, and room temperature beverages. Whenever a test which met the requirement of at least 32 subjects each of whom had tested each sample occured in normal operations, an extended analysis of variance of the following type would be conducted for each background category.

Source of Variation	Degrees of Freedom
Among food samples	x = number of samples minus one
Among background categories	1 = number of categories (2) minus one
Among subjects (within category)	y = number of subjects minus 2
Food samples - background category interaction	x
Food samples - subject interaction	xy

(f) The first objective was to accumulate such sets of analyses for 50 foods, 10 in each of the five categories.

(g) Analyze and evaluate the combined results. Draw final conclusions at this stage if warranted, or suggest additional work, possibly with revisions. The frequency distributions of the background characteristics for each food category would be examined to determine whether categories differed significantly. The importance of any such difference would depend upon whether the interaction between food samples and background category was significant.

2. Results of questionnaire

Questionnaires were completed by 400 individuals, representing slightly less than two-thirds of the total panel.

The distributions of answers are shown in Table 1.1.

It was recommended that four of the factors be omitted:

(a) Military service. Because 107 of the 172 persons who had never been in military service were female, a dichotomy between those who were, or had been, in service and those who had not been would involve a confounding of the sex variable. Thus, it would not be possible to partial out the effect of sex differences in any military service by food sample interaction.

(b) Garrison rations. The reason for omitting this category was the same as for omitting military service. Here 107 of the 220 individuals who had never eaten garrison rations for at least a month were female.

(c) Operational rations. Again, 107 of the 279 persons who had never eaten operational rations for at least three days were female. Hence, this characteristic was also highly confounded with sex differences.

(d) Opinion of taste test program. The responses on this item were largely redundant with those on feelings about taste-test participation.

For the remaining variables, the following dichotomies were established:

(a) Age - Group 1: 39 years of age or younger
- Group 2: 40 years of age or older

(b) Sex - Group 1: Female
- Group 2: Male

(c) State of Origin - Group 1: New England
- Group 2: All other states and countries

(d) Education - Group 1: College graduates
- Group 2: Less than a bachelor's degree

(e) Federal Service - Group 1: Less than 5 years
- Group 2: Five years or more

(f) Feelings about Taste-Test Participation
- Group 1: "Gives a great deal of pleasure"
- Group 2: Any less favorable response

(g) Extent of Test Participation
- Group 1: 20 or fewer
- Group 2: More than 20

Table 1.1, which presents the frequency distributions for each of the 11 factors, characterizes the population of taste-test participants at the Natick Laboratories. The figures are largely self-explanatory. Note, especially, the modes for each factor: 40-49 years of age, males, raised in New England, college graduates, civilians almost evenly divided between veterans and non-veterans, between 10 and 20 years of Federal service, never subsisted on garrison or operational rations, very positive feelings about taste-test participation and the taste-test program, and participation in from 21-40 tests.

This table provides data for evaluating the significance of future changes in the test population.

3. Final outcome

Administrative problems arose which prevented continuation of this phase. Cooperation of Computer Branch personnel and considerable clerical time would have been required, and circumstances did not allow for a smooth follow-through. Hence, the phase was abandoned. Detection of significant food sample by background factor interactions, or proof of the self-selection phenomenon, might not have implied any procedural or methodological changes. Rather, the absence of these effects would have meant greater confidence in generalizing the results to the broader military population. Their presence would have raised some questions of validity. To the extent that food development and research decisions are influenced by Laboratory taste test results, the results of this phase would have provided additional guidance in the determination of how much weight should be given to these data.

Despite the administrative problems, consideration should be given to completion of this study on an in-house basis. The formal analyses should provide much useful information, and, perhaps, additional bonuses in the form of new insights into the functioning of the food evaluation program.

Table 1.1

Frequency Distributions for Background Characteristics

Age	Code No.	Category	N
	1	Under 20 years	17
	2	20-29 years	88*
	3	30-39 years	65
	4	40-49 years	126
	5	50-59 years	85
	6	60 and over	19

*Includes all (23) enlisted military personnel

Sex			
	0	Female	121
	1	Male	279
State of Origin			
	01	New England(Mass,Vt,Conn,RI, NH,Me)	259
	02	East Central(NY,NJ,Pa,Md, Del,Ohio,W.Va,D.C.)	57
	03	Southeast(Ken,Va,Tenn,N.Ca, S.Ca,Miss,Ala,Fla,Ga)	12
	04	Midwest(Ill,Mich,Wisc,Ind)	23
	05	Great Plains(Neb,Iowa,Kans,Mo)	9
	06	South Central(Texas,Okla,Ark,La)	8
	07	Rocky Mts.(Mont,Wyo,Col,Utah, Nev)	1
	08	Northwest(Wash,Ore,Idaho,Alas)	3
	09	Southwest(Calif,NM,Ariz,Hawaii)	3
	10	North Central(No.Da,So.Da,Minn.)	5
	11	Outside U.S.	20
Education			
	1	Grade School	2
	2	High School,not completed	9
	3	High School,graduate	85
	4	College,not completed	80
	5	College,graduated	224

Table 1.1

			N
Military Service	1	Now in service, RA	12
	2	Now in service, US	23
	3	Civilian, in reserves	30
	4	Civilian, Veteran	163
	5	Civilian, never in service	172*
		*Includes 107 females	
Years in Federal Service	1	Less than 1 year	28
	2	One, but less than 2 years	41
	3	2 but less than 5 years	59
	4	5 but less than 10 years	86
	5	10 but less than 20 years	100
	6	20 years or more	86
Garrison Rations	0	No (have never eaten)	220*
	1	Yes	180
		*Includes 107 females	
Operational Rations	0	No	279*
	1	Yes	121
		*Includes 107 females	
Feeling about Taste Test Participation	1	It gives me a great deal of pleasure	194
	2	It gives me a certain amount of pleasure	173
	3	I have no feeling about it one way or the other	25
	4	I feel a slight dislike about participating	5
	5	I definitely dislike participating and do so only out of a sense of duty	3

Table 1.1

		N
Opinion of Taste Test Program	1. It is one of the most important factors in the success of the Army's food research program	237
	2. It helps to an important degree even though it is not crucial to the research program	129
	3. It is moderately useful to the food research program	31
	4. It helps a little, but the time and effort could be used elsewhere to better advantage	3
	5. It is a complete waste of time	0
Number of Taste Tests Participated In	1-20	116
	21-40	193
	41-60	73
	61-80	5
	81-100	4
	100 and over	9

Chapter III. DETERMINATION OF THE LIMITS OF NUMBER OF SAMPLES EVALUATED

The time a judge actually spends in the taste-test booth is only a part of the total time he devotes to testing. Indirect time costs — walking to and from the laboratory, waiting for assignment to a vacant booth, partaking of refreshments — taken together are probably greater than the direct time.

it may be assumed that the more samples an individual can evaluate in one session, the more work the laboratory can produce. For example, doubling the number of samples would cut the subjects' indirect time costs in half. While preparation time could not be reduced by this amount, the time savings could be substantial because samples are prepared in similar ways so that the economies of assembly line practices would be achieved.

It is perhaps even more important that often more straightforward experimental designs can be used instead of the complicated ones such as balanced incomplete blocks. More precise estimates of inter-judge variability and judge-sample interaction can be used in the analyses of variance and the tedium and assumptions inherent in block adjustments can be avoided.

While having a judge evaluate a large number of samples is advantageous, intuitively there is a limit. Certainly, most judges can be induced to cooperate in testing a large number of samples, but does acquiesence bring with it a deterioration in performance? We do not necessarily mean that the deterioration would be due to sensory adaptation or reduction in acuity. Deterioration can also be a function of motivation as reflected in reduced attention, carelessness in rating, general confusion, and other factors resulting in sub-optimal performance.

Much of the earlier literature on this topic was reviewed by Bradley (1954). Bradley noted that from two to eight samples per session have been advocated by various researchers; but there has been no evidence that the recommended number was actually the optimum.

Most of the experiments cited by Bradley dealt with sensory discrimination rather than with preference. Some investigators reported a deterioration in performance with certain types of foods as the number of samples increased, and others found no loss in discrimination with some foods even after presenting as many as 75 samples in one session.

Bradley himself found no differences in hedonic ratings for canned sauerkraut whether served in the first or third vs the fifth or eighth positions in an eight-sample test. The same conclusion held for another experiment involving canned bread and maragarine. Bradley, however, was concerned with the absolute ratings of these products. In most preference tests, the absolute ratings are of secondary importance. The main concern is whether the differences and the direction of differences among competing samples are maintained. Is the algebraic difference, for example, between samples A and B the same regardless of whether the two are served early or late in a series? The replications of the experiment to be reported here concern the testing of one method of increasing the efficiency of taste tests — simply that of extending the number of samples. The purpose of this study was to determine the effect of extended testing on preference differentiation. The basic question is: are the differences in preference among food samples constant whether these foods are served alone or with other samples of the same subclass?

1. Methods

There were three replications of the basic experimental design which differed only in the products tested: soup and gravy base, milk, and syrup.

Judges

In each replication, 40 participants, selected randomly from the taste-test pool, were assigned to one of five experimental conditions. Separate selection was made for each experimental condition, so that a total of 200 participated in each replication.

Experimental Conditions

The experimental conditions varied in the number of samples evaluated. In each session, all samples were evaluated twice by the same judge — once in the first half and again in the second half. In the first condition, judges rated two samples (A and B). In the second, they rated three samples (A, B, and C); in the third, four samples (A, B, C, and D); in the fourth, five samples (A, B, C, D, and E); and in the fifth, six samples (A, B, C, D, E, and F). After rating all samples once, the judges rated them again. The serving orders in each half were balanced to the fullest extent possible. For any one subject, the serving order in the second half was independent of the serving order in the first half. Judges were told only the number of samples to be rated and nothing about duplications. Because each experimental condition was conducted in a separate session, condition and session were confounded.

Normal laboratory practices were followed. The samples were rated on the nine-point hedonic scale.

Selection of Samples

The primary question in this study was whether differences between Samples A and B vary according to the total number of samples evaluated. Accordingly, it seemed advisable to select A and B such that they were not too far apart nor too close together in preference. If the difference were large, then the maximum possible difference in rating might be obtained even in the least suitable condition. If the difference were too small, then even the most sensitive test method might not be able to demonstrate an effect and the negative results would be meaningless.

Hence, the strategy was adopted of pilot testing the series of six samples to be used in the final test and designating the one with the second highest rating as Sample A and the one with the second lowest rating as Sample B. Thus, one sample would be higher than A, two would be between A and B, and one would be lower than B.

This plan was carried out for the experiments with the soup and gravy base and with the milk. For syrup, however, the results of the pretest suggested that this plan be revised. One of the six samples was sorghum, and its pretest rating was exceptionally low (3.77). To have included this item in the test might have induced contrast effects that would reduce discrimination among the other samples. Since sorghum would have been tested only in the sixth condition, the main effects of numbers of samples might have been confounded with contrast effects in some complex manner. Examination of the ratings in the pilot test indicated the advisability of a different approach. The top-rated sample was designated as Sample A and the second-ranked as Sample B since there appeared to be sufficient separation between them in terms of average rating. The lowest rated sample, sorghum, was eliminated, and the third-ranked sample was duplicated as both D & F.

Table 2.1 describes the samples used in each replication, and shows the rank order of preference in the pretest, the serving temperature, the size of the sample, and the serving interval.

Deviation

In the soup and gravy base experiment data were available from only 36 judges for the last experimental condition. How this matter was handled is described later.

2. Results

The results are first reported by replication. For each, the analysis of variance involving Samples A and B across experimental conditions is discussed, then the analyses of variance within each of the five experimental conditions. Finally, for each type of analysis, the consistent and inconsistent findings are summarized.

Soup and Gravy Base

Table 2.2 shows the mean rating of each sample in each experimental condition, by half and by total. The grand totals of all samples within each experimental condition are also shown. These means are based upon N=40, except for the last experimental condition where N=36.

Two separate analyses of variance were run on these data. One used only the data from the first four experimental conditions where there were 40 subjects. The other used the data from all five experimental conditions, but the ratings given by four randomly selected judges were eliminated from each of the first four conditions. The results of the two analyses were essentially the same; hence, the tables and the discussion will be based upon the analysis with 36 subjects.

Table 2.3 shows the effect of the experimental conditions on the sample ratings. For each condition (2-sample through 6-sample) it gives the average of Samples A and B for the first half and second half of the session and the difference between the halves. It also shows the differences between Samples A and B for the two halves of the experiment, and how this difference changes — in direction and size — from one experimental condition to another.

The analysis of variance across all conditions is presented in Table 2.4. Samples A and B differed by 0.61 scale points, which was significant at the 0.1% level. The samples were consistently rated higher in the first half of the session than in the second; the average across all five conditions was 0.44 scale points. The main effect of experimental condition was significant (1% level). A considerable part of this effect is probably attributable to the 6-sample condition, where the samples were rated particularly low (See Table 2.2). The main effect of condition was not significant in the analysis which excluded the 6-sample experiment. It may be noted that the average ratings for a given sample tended to decrease as the number of samples increased (Table 2.2).

No interaction involving sample, half, or experimental condition was statistically significant. We would conclude, therefore, that although ratings tended to be lower in the second half than in the first, there is no evidence that the differences among samples were affected either by the number of samples or by whether they were presented in the first half or in the second half of the session.

A separate analysis of variance of the ratings for Samples A and B was performed for each experimental condition. A summary of these analyses appear in Table 2.5. The difference between A (6.91) and B (6.66) was not significant in the 2-sample condition. With this exception the ratings always differed significantly ($p < .001$). The main effect of experimental half was always significant, at the .001 level for the last four conditions, and at the .05 level for the 2-sample condition. In the 3-sample condition, the interaction of sample and half was statistically significant ($p < .05$). The range of ratings was .55 for the first half, but 1.40 for the second half; however, this increased level of differentiation among samples as a function of the half session in which they were presented was not evident in the other four conditions. No other main effect was statistically significant.

There was no evidence that the number of samples had a consistent effect upon differences in preference ratings between the two selected samples. While the samples in the second half were generally rated lower than those in the first half, in the analysis of variance involving A and B only, and in only one of the five analyses of variance of all samples within condition, were the differences among samples shown to depend upon whether they appeared in the first or the second half.

Milk

The analyses here were the same as those for the soup and gravy base. Table 2.6 lists, by experimental condition, the mean ratings of each sample in each nalf and for both halves combined.

Table 2.7 is a simplification of Table 2.6, but concerns only Sample A and B. The averages of the two samples in the first and the second halves, and the differences between the halves, are shown for each condition and over all conditions. Also given are the mean differences between Samples A and B for each condition and each half, as well as the differences between these values.

The analysis of variance shown in Table 2.8 indicates that the two samples were significantly different in preference. The grand average of Sample A was 6.69 and that of sample B was 5.61. The difference of 1.08 was larger than had been originally anticipated. The absence of a significant interaction between sample and half means that the difference between the samples in the first half (.97) was not significantly smaller than their difference in the second half (1.21). As in the preceding replication, the ratings in the first half were significantly ($p <.05$) higher than those in the second half — by an average of .19 scale points.

None of the interactions were significant, and there is no evidence that the number of samples decreased or increased the difference between Samples A and B, or had any other significant effect upon the level of ratings.

The figures in Table 2.7 would seem to indicate otherwise. The "second-half minus first-half" difference appears to drop as the number of samples increases, from 0.60 for the 2-sample condition successively to .49, .08, .08, and -.08. But the three-factor interaction was not significant, so this apparent trend has no statistical support.

An analysis of variance was run for each experimental condition separately (Table 2.9). Again, the main effect of sample was in each case highly significant ($p <.001$). The only other significant source of variation in any analysis was the main effect of half in the six-sample condition ($p <.01$), where the difference between the first and second half was .35. In the 2-sample through the 5-sample conditions, the differences between halves were, consecutively: .35, .02, .18, and .24. (See totals in Table 2.6). Once more, there is no evidence that the number of samples affected the differences between Samples A and B.

Syrup

The means of each sample in each experimental condition, by first and second half and by total, are listed in Table 2.10. The analyses of variance confirm certain features which are apparent in this table. Reference to Table 2.11, which shows the means of, and the differences between Samples A and B for each half, clarify the interpretation of the analysis of variance in Table 2.12.

The two samples were significantly different in preference (p .001). In the pretest, Sample A (Finest) had rated 0.57 scale points higher than Sample B (Long Cabin): however, there was a definite reversal in the main experiment. Considering the average across all five conditions, Sample B rated 0.34 scale points higher and there were only two instances where Sample A was higher in any half of any condition. This gross failure in prediction cannot be explained on the basis of available information. It was fortunate that the samples proved to be different, otherwise this replication would have been wasted effort.

Just as in the first two replications, the effect of first vs second half was significant ($p < .05$), with the second-half rating lower. The effect of experimental condition was also significant ($p < .05$). The means of A and B ranged from 6.90 (2-sample) to 6.00 (6-sample): however, the decrease was not monotonic.

The only other significant effect ($p < .05$) was the interaction between half and condition. The right-hand portion of Table 2.11 illustrates this, but there is no clear reason why it occurred. In the first half the difference between the samples was highest for the 3-sample and 5-sample conditions; in the second half, it was highest for the 2-sample and 5-sample conditions and almost zero for the 4-sample and 6-sample conditions. There is no linear trend for the difference to vary according to the number of samples tested, whether one considers each half separately or both combined.

The absence of an interaction between samples and conditions, or between samples and half, is consistent with the results from the other two replications. Thus, the level of ratings might be affected by whether the samples appeared in the first or second half, but, on an overall basis, the conclusion that Sample B is preferred to Sample A does not depend upon the half in which they were tested or upon the number of samples with which they were evaluated, or upon the interaction of the two variables.

The separate analyses of variance which were performed for each experimental condition are summarized in Table 2.13. Once again, significant differences among samples are clearly demonstrated. The level of significance was .05 for the 2-sample condition and .001 for the other four conditions. Similarly, in each case the ratings in the second half were significantly lower than the ratings on the first half.

In the 4-sample and in the 5-sample conditions, the interactions between sample and half were significant ($p < .001$, and $< .01$ respectively). In the former condition, pure maple syrup dropped by 1.32 scale points from the first to the second half, and the Government Standard syrup dropped by 1.20 scale points. However, the mean rating of Finest was only .27 scale points lower on the second half, and the mean rating of Log Cabin only .08 scale points lower. Similarly, in the 5-sample condition, the pure maple and Government Standard syrups dropped 1.27 and 1.37 scale points, respectively, in the second half; but Finest and Log Cabin dropped only .05 and .10 scale points. Thus, the nature of the interaction between sample and half was nearly the same in each of the two conditions.

All four samples appearing in the 4-sample and 5-sample conditions also were present in the 6-sample condition, yet the sample-by-half interaction was not significant. Why not? It may be noted (Table 2.10) that the mean ratings for the pure maple and the Government Standard syrups were initially lower in the first half of the 6-sample condition than in the 4-sample and 5-sample conditions. The two samples did drop in the second half of the 6-sample condition by at least .65 scale points, but for some reason Log Cabin also dropped by a larger amount, .45 scale points, than in the preceding two conditions. Thus, the absence of a sample-by-half interaction in the 6-sample condition may have been due to abnormally low first-half ratings for the pure maple and the Government Standard syrups and/or to the abnormally large decrease for Log Cabin. Either or both of these effects would work in the direction of supporting the null hypothesis for the sample-by-half interaction.

Even in the two conditions where the sample-by-half interaction was significant, the rank order of prefer nce within half remained the same: Log Cabin always had the highest mean rating, Finest the second highest, Government Standard next, and the pure maple syrup the lowest. The range of ratings was somewhat higher in the second half than in the first half: 2.37 vs 1.30 for the 4-sample condition and 3.37 vs 2.20 for the 5-sample condition. There is no evidence that the differences among samples are attenuated in the second half, or vary among experimental conditions.

One more point deserves mention. Note the mean values for samples D and F in the 6-sample condition. These samples were identical (Government Standard). The ratings in the first half were 5.55 and 5.47, and in the second half they were 4.90 and 4.80. The reliability seems good.

3. Discussion

All three replications yielded similar interpretations. First, the ratings in the second half were significantly lower than in the first half. This point is inconsequential when considering one major purpose of taste-tests, which is to determine differences among samples rather than to establish levels of rating. However, some users of the data might be pleased or dismayed at the lower ratings in the second half and might have to shift their frame of reference when using such ratings as absolutes.

Second, within each replication the samples differed sufficiently to allow the effects of other variables to come into play. An unanswered question is whether the differences were, in fact, too great so that they obscured the effects of these other variables. This may have occurred for the second replication, milk, where the overall mean difference between Sample A and Sample B was 1.09 scale points (Table 2.7); but the difference of .61 for soup and gravy base (Table 2.3) and of .34 for syrup (Table 2.11) approach the specifications for this experiment. But, because the conclusions corresponded so well among the three replications, we do not believe that the large difference between milk Samples A and B constitutes a serious problem.

The fact that the mean ratings varied among experimental conditions in the first and third replications does not in itself mean very much. These effects could be caused by the effects of the other samples evaluated with A and B or by the psychological effects on the judges of having to rate varying numbers of samples.

The crucial source of variation is the interaction between experimental condition and specific samples. In none of the three analyses of variance involving only Samples A and B was this source significant.

In the analyses of variance of the individual conditions, in only one condition of the soup and gravy base replication and in two conditions of the syrup replication, was the interaction between sample and half significant. In each of these cases, the rank orders of preference in the two halves were identical. If anything, the ratings were spread out more in the second half than in the first half. Only one other source of variation in any analysis of variance was significant — the interaction between condition and half in the last replication -- and the nature of the significance has no clear importance.

Thus, there is no evidence that increasing the number of samples up to 12 would change the relative differences in ratings among food samples. If there is some upper limit, we have not attained it. These negative results are meaningful and important since they give one confidence in conducting taste tests involving a greater number of samples than the typical three, four, or five. However, we do not know the effect of lengthened tests upon the panel population. Perhaps most participants would not object to an occasional ten - or twelve - sample test; but if such tests became common, then some might be induced to withdraw. Perhaps some people especially welcome frequent and short tests as breaks from their everyday activities, and might dislike less frequent and longer ones. Even so, an occasional longer test might productively be used. There is no contrary experimental evidence.

The Natick laboratory noted that certain difficulties were encountered in conducting the more extended sessions. For example, counter space was limited to accommodate the larger number of samples, the routines of preparing and serving samples, calling subjects, etc. created a peak work-load beyond the capacity of the regular operators to handle, and it was difficult to fit a session (experimental condition) into the customary half-day test period. Further, individual subjects were often delayed in waiting for a free test booth. That practical problems such as these might arise was anticipated. Their presence does not obviate the value of the basic conclusions; however, they do indicate the need for special planning and special arrangements.

The data suggest several other problems worthy of study. A few significant interactions between sample and half were noted. It would seem that some samples are affected by certain others with which they are served, such that in the second half they drop disproportionately. This phenomenon, which did not always appear, might be a manifestation of the contrast and convergence effects which are discussed in the next section. Certain samples of a product might achieve a fairly high average rating the first time, but drop when tested again, because some judges do not become aware of their deficiencies until they have had intervening experience with good quality products. If some samples are disproportionately affected when repeatedly evaluated by the same person, then one would hypothesize that they would be more subject to monotony effects if they became standard items of issue than samples which showed only the normal loss in preference in the extended test situation. The reason is that the more often a susceptible food is served, the greater the opportunity for deficiencies to be noticed. For example, if firmer evidence were available that Karo and pure maple syrup are nearly equal in preference, one might hypothesize that preference for the latter would decline more sharply than preference for the former, assuming an equal rate of use.

In view of the absence of evidence that the number of samples affects the differences in ratings among food samples, there seems to be little reason for further exploring procedural methods, such as those suggested at the onset of this contract, for increasing efficiency. If efficiency does not decrease with even up to twelve samples, then introducing such devices as rest breaks or increasing the variety of different food types evaluated in a session could serve no practical purpose.

If it becomes customary for a large number of samples to be evaluated in a session, then occasional checks should be made to insure that performance is not deteriorating because of motivational factors. Checks should also be made on the panel drop-out rates.

Of course, it may be unwarranted to generalize from three foods, particularly when neither a meat or a vegetable was included. Additional replications of this experiment, conducted and analyzed in-house, would increase confidence in the main conclusion and in the implications of the conclusion. For example, if different results were found with meats, then one could set better limits on the applicability of the present results.

Some comments on the experimental design may be helpful. In this study Samples C through F were only of secondary importance. They were "filler" samples needed to achieve the experimental specifications. However, it was desirable that they differ qualitatively and in preference value since each experimental condition was analyzed separately; and each analysis contributed to the conclusions dealing with the effect of test length upon differences in preference among samples. One experimental alternative, if interest had been exclusively focused upon the differences in preference between Samples A and B, would have been to assign at random any one of Samples C through F as the third sample in the 3-sample condition; so that these third samples would differ among judges. In this way, A and B would be tested approximately an equal number of times with C, with D, with E, and with F. A comparable arrangement could have been made in the 4-sample and 5-sample conditions. Then the results would not have depended as heavily upon the individual sample, or samples, which accompanied A and B. However, this method would have precluded the analysis of the individual experimental conditions and would probably have increased experimental error. Considering all factors, the method used seemed preferable.

Table 2.1. Description of Samples Used in Each Replication

Product	Lots/Formulations	Rank Order Pre-test	Serving Temperature (°F.)	Size of Serving	Serving Interval
Soup and Gravy Base	A (Prepared by the	2	150-155	2 oz.	30 sec.
	B Nestle Company)	5			
	C	3			
	D	4			
	E	1			
	F	6			
Milk	A 92% fresh 8% reconstituted*	2			
	B 68% fresh 32% reconstituted	5	Room Temp.	1 oz.	30 sec.
	C 84% fresh 16% reconstituted	3			
	D 76% fresh 24% reconstituted	4			
	E 100% fresh 0% reconstituted	1			
	F 60% fresh 40% reconstituted	6			
Syrup	A Finest	1			
	B Log Cabin	2			
	C Pure Maple	4	115-120	½ oz.	60 sec.
	D Government Std.	3			
	E Karo	5			
	F Government Std.	3			

* Reconstituted: 33% Pet mi!k, 67% water

Table 2.2. Soup and Gravy Base. Mean Ratings of Samples in Each Experimental condition (N=36, last condition, 40 all others)

	Experimental Condition				
	2 Sample	3 Sample	4 Sample	5 Sample	6 Sample
Sample A					
1st half	7.13	6.93	6.68	6.83	6.58
2nd half	6.70	6.98	6.45	6.33	5.81
Total	6.91	6.95	6.56	6.58	6.19
Sample B					
1st half	6.85	6.38	6.28	6.13	5.67
2nd half	6.48	5.58	5.88	5.48	5.14
Total	6.66	5.98	6.08	5.80	5.40
Sample C					
1st half		6.85	6.90	6.93	6.28
2nd half		6.05	6.23	6.63	6.25
Total		6.45	6.56	6.78	6.26
Sample D					
1st half			6.80	7.33	6.86
2nd half			6.80	6.68	6.53
Total			6.80	7.00	6.69
Sample E					
1st half				6.88	6.83
2nd half				6.63	6.06
Total				6.75	6.44
Sample F					
1st half					4.55
2nd half					3.81
Total					4.18
Total					
1st half	6.99	6.72	6.66	6.82	6.13
2nd half	6.59	6.20	6.34	6.35	5.60
Total	6.79	6.46	6.50	6.58	5.86

Table 2.3. Soup and Gravy Base. Effect of Experimental Condition on Ratings of the Critical Samples (N=36 per condition*)

Experimental Condition	Average of Sample (A+B)			Difference between samples (A-B)		
	1st half	2nd half	Difference 1st - 2nd	1st half	2nd half	Difference 2nd - 1st
2-Sample	7.02	6.56	.46	.19	.23	.04
3-Sample	6.72	6.34	.38	.55	1.40	.85
4-Sample	6.44	6.24	.20	.50	.47	-.03
5-Sample	6.42	5.90	.52	.67	.69	.02
6-Sample	6.12	5.48	.64	.91	.67	-.24
Total	6.54	6.10	.44	.55	.66	.11
Average of 1st & 2nd halves	6.32			.61		

* Ratings given by four randomly selected judges were eliminated from each of the first four conditions.

Table 2.4. Soup and Gravy Base. Analysis of Variance For Critical Samples (A and B) Across Experimental Conditions (N=36 per condition)

Source of Variation	df	ms	F
A-Sample	1	66.61	35.44***
B-Half	1	35.11	30.27***
C-Condition	4	20.16	3.65**
A x B	1	.51	#
A x C	4	2.57	1.37
B x C	4	.98	#
A x B x C	4	1.28	1.58
D-Judge (within C)	175	5.53	
A x D	175	1.88	
B x D	175	1.16	
A x B x D	175	.81	

** Significant at the 1% level.
*** " " " .1% " .

# F-ratio less than 1.00.

Testing of effects:

A	tested	against	A x D
A x C	"	"	A x D
B	"	"	B x D
B x C	"	"	B x D
C	"	"	D
A x B x C	"	"	A x B x D
A x B	"	"	A x B x D

Table 2.5. Soup and Gravy Base. Summary of Analyses of Variance for Each Experimental Condition (N=36 last condition, 40 all others)

Source of Variation	2-Sample df	2-Sample F or (ms)	3-Sample df	3-Sample F or (ms)	4-Sample df	4-Sample F or (ms)
A-Sample	1	1.37	2	10.65***	3	5.76***
B-Half	1	6.40*	1	14.17***	1	8.05***
D-Judge	39	(4.28)	39	(6.66)	39	(11.38)
A x B	1	#	2	3.68*	3	1.49
A x D	39	(1.83)	78	(1.79)	117	(1.29)
B x D	39	(1.44)	39	(1.13)	39	(1.05)
A x B x D	39	(1.64)	78	(1.31)	117	(1.08)

Source of Variation	5-Sample df	5-Sample F or (ms)	6-Sample df	6-Sample F or (ms)
A-Sample	4	9.26***	5	22.41***
B-Half	1	20.64***	1	18.44***
D-Judge	39	(9.46)	35	(11.25)
A x B	4	#	5	1.31
A x D	156	(1.82)	175	(2.79)
B x D	39	(1.07)	35	(1.66)
A x B x D	156	(1.22)	175	(1.27)

* Significant at the 5% level.
*** " " " .1% " .

# F ratio less than 1.00.

Testing of effects:

A tested against A x D
B " " B x D
A x B " " A x B x D

Table 2.6. Milk. Mean Ratings of Samples In Each Experimental Condition (N=40 each condition)

	Experimental Condition				
	2	3	4	5	6
	Sample	Sample	Sample	Sample	Sample
Sample A					
1st half	6.78	6.65	6.75	6.90	6.58
2nd half	6.73	6.75	6.75	6.68	6.38
Total	6.75	6.70	6.75	6.79	6.48
Sample B					
1st half	5.95	5.67	6.08	5.75	5.35
2nd half	5.30	5.28	6.00	5.45	5.23
Total	5.63	5.48	6.04	5.60	5.29
Sample C					
1st half		6.23	6.63	6.30	6.08
2nd half		6.45	6.33	6.15	5.82
Total		6.34	6.48	6.23	5.95
Sample D					
1st half			6.45	5.90	5.85
2nd half			6.13	5.60	5.55
Total			6.29	5.75	5.70
Sample E					
1st half				6.90	6.75
2nd half				6.58	6.35
Total				6.74	6.55
Sample F					
1st half					5.45
2nd half					4.63
Total					5.04
Total					
1st half	6.36	6.18	6.48	6.35	6.01
2nd half	6.01	6.16	6.30	6.09	5.66
Total	6.19	6.17	6.39	6.22	5.83

Table 2.7. Milk. Effect of Experimental Conditions on Ratings of the Critical Samples (N=40, each condition)

	Average of Samples (A+B/2)			Difference Between Samples (A-B)		
	1st half	2nd half	Difference 1st - 2nd	1st half	2nd half	Difference 2nd - 1st
Experimental Condition						
2-Sample	6.36	6.01	.35	.83	1.43	.60
3-Sample	6.16	6.02	.14	.98	1.47	.49
4-Sample	6.42	6.38	.04	.67	.75	.08
5-Sample	6.32	6.06	.26	1.15	1.23	.08
6-Sample	5.96	5.80	.16	1.23	1.15	-.08
Total	6.24	6.05	.19	.97	1.21	.24
Average of 1st & 2nd halves	6.14			1.09		

Table 2.8. Milk. Analysis of Variance For Critical Samples (A and B) Across Experimental Conditions (N=40, each condition)

Source of Variation	df	ms	F
A-Sample	1	236.53	123.19***
B-Half	1	7.41	6.18*
C-Condition	4	5.56	#
A x B	1	2.77	2.72
A x C	4	1.81	#
B x C	4	.56	#
A x B x C	4	.88	#
D-Judge (within C)	195	6.82	
A x D	195	1.92	
B x D	195	1.20	
A x B x D	195	1.02	

* Significant at the 5% level.
*** " " " .1% " .

# F-ratio less than 1.00.

Testing of effects:

A	tested	against	A x D
A x C	"	"	A x D
B	"	"	B x D
B x C	"	"	B x D
C	"	"	D
A x B x C	"	"	A x B x D
A x B	"	"	A x B x D

Table 2.9. Milk. Summary of Analyses of Variance for Each Experimental Condition (N=40, each condition)

	2-Sample		3-Sample		4-Sample	
Source of Variation	df	F or (ms)	df	F or (ms)	df	F or (ms)
A-Sample	1	36.95***	2	18.10***	3	5.10***
B-Half	1	3.02	1	#	1	1.48
D-Judge	39	(8.37)	39	(7.43)	39	(15.50)
A x B	1	3.31	2	2.73	3	#
A x D	39	(1.37)	78	(1.75)	117	(1.42)
B x D	39	(1.62)	39	(.96)	39	(1.66)
A x B x D	39	(1.09)	78	(.80)	117	(.84)

	5-Sample		6-Sample	
	df	F or (ms)	df	F or (ms)
A-Sample	4	13.36***	5	18.53***
B-Half	1	3.13	1	7.35**
D-Judge	39	(13.41)	39	(17.84)
A x B	4	#	5	1.36
A x D	156	(1.79)	195	(1.63)
B x D	39	(2.16)	39	(2.00)
A x B x D	156	(1.33)	195	(.92)

** Significant at the 1% level.
*** " " " .1% " .

# F ratio less than 1.00.

Testing of effects:

A tested against A x D
B " " B x D
A x B " " A x B x D

Table 2.10. Syrup. Mean Ratings of Samples in Each Experimental Condition (N=40, each condition)

	Experimental Condition				
	2 Sample	3 Sample	4 Sample	5 Sample	6 Sample
Sample A					
1st half	7.03	6.75	6.60	6.28	5.90
2nd half	6.25	6.38	6.33	6.23	5.90
Total	6.64	6.56	6.46	6.25	5.90
Sample B					
1st half	7.38	7.38	6.43	6.95	6.33
2nd half	6.93	6.60	6.35	6.85	5.88
Total	7.15	6.99	6.39	6.90	6.10
Sample C					
1st half		5.45	5.30	4.75	3.55
2nd half		4.70	3.98	3.48	2.88
Total		5.08	4.64	4.11	3.21
Sample D					
1st half			5.93	5.70	5.55
2nd half			4.73	4.33	4.90
Total			5.33	5.01	5.23
Sample E					
1st half				4.88	5.23
2nd half				4.35	4.80
Total				4.61	5.01
Sample F					
1st half					5.47
2nd half					4.80
Total					5.14
Total					
1st half	7.20	6.53	6.06	5.71	5.34
2nd half	6.59	5.89	5.34	5.05	4.86
Total	6.90	6.21	5.70	5.38	5.10

Table 2.11. Syrup. Effect of Experimental Conditions on Ratings of the Critical Samples (N=40, each condition)

Experimental Condition	Average of Samples (A+B/2) 1st half	2nd half	Difference 1st - 2nd	Difference Between Samples (B-A) 1st half	2nd half	Difference 2nd - 1st
2-Sample	7.20	6.59	.61	.35	.68	.33
3-Sample	7.06	6.49	.57	.63	.22	-.41
4-Sample	6.52	6.34	.18	-.17	.02	.19
5-Sample	6.62	6.54	.08	.67	.62	-.05
6-Sample	6.12	5.89	.23	.43	-.02	-.45
Total	6.70	6.37	.33	.38	.30	-.08
Average of 1st & 2nd halves	6.54			.34		

Table 2.12. Syrup, Analysis of Variance For Critical Samples (A and B) Across Experimental Conditions (N=40, each condition)

Source of Variation	df	ms	F
A-Sample	1	23.46	13.72***
B-Half	1	11.76	6.53*
C-Condition	4	19.45	3.34*
A x B	1	.56	#
A x C	4	3.25	1.90
B x C	4	4.99	2.77*
A x B x C	4	1.14	1.24
D-Judge (within C)	195	5.83	
A x D	195	1.71	
B x	195	1.80	
A x B x D	195	.92	

* Significant at the 5% level.
*** " " " .1% " .

# F-ratio less than 1.00.

Testing of effects:

A	tested	against	A x D
A x C	"	"	A x D
B	"	"	B x D
B x C	"	"	B x D
C	"	"	D
A x B x C	"	"	A x B x D
A x B	"	"	A x B x D

Table 2.13. Syrup. Summary of Analyses of Variances For Each Experimental Condition (N=40, each condition)

	2-Sample		3-Sample		4-Sample		5-Sample		6-Sample	
Source of variation	df	F or (ms)	df	F or (ms)	df	F or (ms)	df	F or (ms)	df	F or (ms)
A-Sample	1	5.00*	2	30.22***	3	15.11***	4	20.84***	5	19.14***
B-Half	1	4.52*	1	11.26**	1	11.27**	1	23.53***	1	9.40**
D-Judge	39	(3.32)	39	(5.89)	39	(11.24)	39	(14.63)	39	(11.07)
A x B	1	3.15	2	#	3	9.44***	4	4.58**	5	#
A x D	39	(2.10)	78	(2.67)	117	(4.10)	156	(5.18)	195	(4.37)
B x D	39	(1.80)	39	(2.14)	39	(2.55)	39	(1.88)	39	(2.93)
A x B x D	39	(.34)	78	(1.29)	117	(1.30)	156	(1.74)	195	(1.61)

* Significant at the 5% level.
** " " " 1% " .
*** " " " .1% " .

F ratio less than 1.00.

Testing of effects:

A tested against A x D
B " " B x D
A x B " " A x B x D

Chapter IV. Further Study of Contrast and Convergence Effects

Much of the food development effort is devoted to better utilization of existing foods. New processing methods are intended to confer upon these foods characteristics which enhance their usefulness in military situat'ons. For example, dehydration can reduce weight and increase storage life, irradiation can eliminate the need for refrigeration or minimize certain types of spoilage.

One of the criteria for evaluating these processes is the extent to which human beings like or dislike the products. For each major method, such as irradiation, there may be an almost infinite number of specific processing variations and combinations of variations. Within limits imposed by cost and manufacturing capability, a major goal is to produce foods which, even if they do not taste the same as the fresh equivalents, are as close to them as possible physically and chemically and are liked equally as well. Conditions and duration of storage are other factors which may interact with processing variables. Assessment of the effects of all, or most, of these is usually necessary to insure that the product as eaten meets preference standards.

Taste tests often serve the dual function of testing processed foods against fresh equivalents and comparing different processing variables or different values of a single processing variable among themselves. The first function involves comparing a control with a series of experimental samples; for example, a fresh food might be compared with the same product subjected to varying degrees of irradiation. The second function, in the case of irradiation, would be represented by comparisons only among the treated samples to determine whether any of the levels caused less flavor damage.

It has been pointed out (Kamenetzky, 1959; Eindhoven, Peryam, Heiligman, and Hamman, 1964) that these purposes might be incompatible. For example, a fresh control might induce contrast effects with the irradiated samples by making the subject aware of negative characteristics that he would otherwise ignore. Hence, the ratings of the irradiated foods following the fresh sample would be depressed. The difference between the control and irradiated samples as a group would be increased but the difference among the irradiated samples themselves might be reduced. Conversely, when a fresh (untreated) sample is tasted in the absence of any other, it may be rated high; however, at least for some foods, negative characteristics are inherent in both the fresh and irradiated forms, which differ only in degree. Tasting an irradiated sample first may call particular attention to such negative characteristics in the fresh control If so, then the fresh food may be rated lower so that the difference between the control and the irradiated samples as a group will be less. This phenomenon is known as convergence effect.

This study is not concerned with demonstrating the existence of contrast and convergence effects *per se*. Rather, they serve as the theoretical basis for an experimental test of whether or not the two sensory evaluation goals are compatible. These phenomena suggest that they may not be. For example, if contrast effect were dominating, then irradiated samples served after the fresh control would likely be "squeezed" together because their negative qualities would be emphasized more than their positive ones. If convergence were operative, then the fresh control following irradiated samples would be rated more like the irradiated samples and the fresh-irradiated difference would be minimized.

This study was oriented toward obtaining information about the compatibility of the two purposes of taste testing. Specifically, the objectives were to determine: (a) whether the presence of a fresh control has an effect upon differences in ratings among irradiated samples, and (b) whether the placement of a fresh control in a series has any effect upon differences among irradiated samples or upon the difference between the control and the irradiated samples as a group.

No direct test of the contrast-convergence theory was intended.

1. Methods

Four replications of this study were conducted. Replication III was considered suspect because of the nonuniformity of the samples served in various sessions; however, all available data are reported as a matter of record. The experimental design and procedures were similar in all four replications. Variations might have occurred because the pool of available subjects changed from time to time, and some deviations might have occurred as a function of different laboratory conditions and diurnal factors.

Subjects

A total of 216 people were separately and randomly drawn for each replication; and 36 were randomly assigned to each of six experimental conditions.

Samples

Let: X, Y, and Z designate three irradiated samples;

FC designate the fresh (unirradiated) control;

IC designate the variable, irradiated control. This means that for one-third of the testers the control sample is X, for another third it is Y, and for the remaining third it is Z.

The purpose of the variable control was to provide a basis for comparisons involving the fresh control by maintaining equivalent positions and numbers of samples.

Experimental Conditions

1. Samples served: FC, X Y, & Z. FC was always served first and then X, Y, and Z in balanced order.

2. Samples served: IC, X, Y & Z. IC was always served first and then X, Y, Z in balanced order.

3. Samples served: X, Y, Z, & FC. FC was always served last, and X, Y, and Z preceded it in balanced order.

4. Samples served: X, Y, Z & IC. IC was always served last, and X, Y, and Z preceded it in balanced order.

5. Samples served: X, Y, Z, and FC. All four samples were served in balanced order. This is the usual design used at the Natick laboratories.

6. Samples served: X, Y, Z and IC. All four samples were served in balanced order.

(In the first four conditions above, FC and IC are underlined to show that the serving order was fixed as either first or last).

Replications

The experimental design was repeated four times, twice with roast beef and twice with ham. Table 3.1 shows the foods, the processing variables, and their levels for each replication. Analysis of the first two replications revealed certain unexplainable inconsistencies in the results. In these replications, experimental condition was confounded with session; that is, only one experimental condition was run at each session. It was possible that the population of available subjects was not the same among sessions, that the physical characteristics of the samples differed from session to session, or that unknown conditions (e.g., weather) may have changed to cause these odd effects. Accordingly, it was decided that for the succeeding replications, each of the six experimental conditions would be represented in each session. The total testing time remained constant.

For the third replication problems were encountered in maintaining physical uniformity of samples from session to session. In fact, the experiment had to be interrupted because of certain difficulties in processing and physical testing; and the samples in the latter part of the test might not have been equivalent to those served in the first part.

Procedure

The standard procedures for the sensory evaluation of foods were used. The samples were served one at a time according to the specifications of the experimental group to which a judge was assigned. After tasting and rating one sample, the subject rinsed his mouth with charcoal-filtered distilled water. Thirty seconds elasped from the time he finished one sample to the time he received the succeeding one. The standard nine-point hedonic scale was used.

2. Results

Tables 3.2-I through 3.2-IV give, for Replications I-IV, respectively, the means for each sample in each experimental condition. Various averages are also shown: The average of the experimental samples, averages by position of control and by type of control, and the grand average. As will be shown later, the differences between experimental samples differed significantly, except in Replication III.

In Replication I (Table 3.2-I), the overall rank order of the experimental samples generally agreed with the hypothesized effect of the temperature variable in that the sample irradiated at ambient temperature was significantly lower than the other two; however, the -80°C. sample ranked higher than the -185°C. sample (5.98 vs 5.90). The rank order of these two samples was different from the overall rank for the two conditions where the control was served first and where the irradiated control was served in the balanced order.

In Replication II, (Table 3.2-II), there was again general agreement with the expected effect of the physical variable. The 3.0 Mrad sample rated highest (6.28), but the 4.5 Mrad and 6.0 Mrad samples were about the same (5.82 and 5.94, respectively). There were many inversions in the rank positions of the latter two samples; however the 3.0 Mrad sample always ranked highest except when the fresh control was served first.

In Replication III, (Table 3.2-III), inconsistencies were the rule rather than the exception. The overall rank order of the experimental samples agreed with the expected effect of the experimental treatments, with the ambient temperature lowest and -40°C. highest, although the range was small 6.05-6.30. When the conditions are considered separately, however, it is seen that none of the experimental samples was consistently the highest or lowest rated.

Replication IV (Table 3.2-IV) resembled Replication II in that the 3.5 Mrad sample was significantly preferred to the 4.5 Mrad or the 6.0 Mrad sample, which in turn did not differ significantly from one another. The 3.5 Mrad sample was the highest rated in each of the six experimental conditions, but sometimes the 4.5 Mrad sample was rated higher than the 6.0 Mrad one, and sometimes the opposite occured.

Analysis of Variance

Table 3.3 shows the results of an analysis of variance of each replication. The interpretation of these analyses will be facilitated by reference to Table 3.4 which shows the algebraic differences in mean ratings between samples related to the experimental variables. We will first describe the outcome for each source of variation, then each replication in turn will be more carefully examined.

(A) Position of Control Sample

There is no evidence that the position of the control sample had an effect upon the overall level of ratings in any of the four replications. The ranges of the averages of the four samples according to the position of the control were .35, .10, .31, and .04 for the four replications, respectively. (See Table 3.2).

(B) Type of Control

Only for Replication II did the type of control have a significant effect upon the overall level of rating. The lack of significant effects in the other three replications is not surprising since three of the four samples were identical across experimental conditions within each replication. This fact would "work" in favor of the null hypothesis, in that the rating of the fresh control would be washed out by the ratings of the other three samples.

(A) X (B)-Type Interaction

None of the four interactions between position of control and type of control was significant.

(C_1) Among Experimental Samples

The presence of significant differences in preference among the samples was a necessary condition in these experiments. The different temperatures at which the foods were irradiated and the different dosage levels were intended to result in appreciable differences in hedonic ratings. For all replications, except Replication III, this intention was fulfilled at very high levels of statistical significance. For Replication III, the range of differences was only .25 scale points, although the order of preference of the three samples corresponded to a priori expectations. If only for this reason, the data from Replication III should be discarded.

(C_2) Between Experimental and Control Samples

For each replication, the average rating of the control samples substantially exceeded the average rating of the experimental samples, even though the differences were attentuated in that in half the cases the control (e.g., irradiated control) was identical to the experimental samples. (See the interactions involving (C_2).

(D) Among judges

This source of variation was significant compared to the interaction of judge and sample. No special importance is attributed to this occurance.

(A) X (C_1)

The interaction between position of control and the difference between experimental samples failed to be statistically significant in any replication.

(A) X (C_2)

This effect was significant except in Replication I. The difference between the control and the average of the experimental samples was higher when the control was served first than when it was served last or in balanced order. For example, in Replication II, the differences for the first, last and balanced positions were .93, .27, and .10 respectively. In Replication III, the differences were 1.22, .20, and .49. In Replication IV, they were 1.19, .16, and .50. The differences fell in the same order in Replication I as they did in II and IV, but failed to reach the 5% level of significance. (See also discussion of (A) X (B) X (C_2).

(B) X (C_1)

In no replication was this effect statistically significant. Thus, there is no evidence that the differences among the experimental samples were any different whether the contrc' was fresh or irradiated. This conclusion is one of the most notable in this study.

(B) X (C_2)

As expected, this interaction was significant in all four replications. Indeed, any other result would have occasioned surprise and even consternation. The significance of these interactions shows that the difference between the experimental and control samples in each replication depended in large part on whether the control was fresh or irradiated. Obviously, one would not expect to find a significant difference between an irradiated control and the experimental samples. In Replication I, the differences between control and the average of the experimental samples were .89 for the fresh control and .07 for the irradiated control. In Replication II, the differences were .82 and .05; in Replication III, 1.20 and .07; and in Replication IV, .97 and .25.

(A) X (B) X (C_1)

This source of variation was significant in the first two replications; but not in the second two. The fact that the nature, or cause, of the significant interactions was not the same in Replication I and Replication II is somewhat disturbing. The variation (Table 3.4) did not follow the same pattern. For example. in Replication I, the largest differences between the theoretically best and theoretically poorest experimental samples occurred for the fresh control served either in the first position or in the balanced position; for the irradiated control the largest difference was in the served last" position. These same relationships held also in regard to the maximum range between experimental samples. However, in Replication II, the largest difference between the experimental samples occurred when the fresh control was served last. In fact, when the fresh control was served first, the theoretically poorest sample was rated higher than the theoretically best. Similarly, the differences among experimental samples was greatest for the irradiated control when it was served first. (This effect might be attributable to position effects, but it is not consistent with the effects observed in Replication I).

The inconsistent directions of these effects make any generalization from these data alone questionable.

(A) X (B) X (C_2)

This effect was significant in all cases except for Replication III, which is of questionable validity anyway. The significance levels were lower than for most other effects. The interpretation of this ef ect seems simple. Referring to Table 3.4, we see that the differences between the control and the average of the experimental samples were in each case highest when the control sample was fresh and was served first. The difference between the irradiated control and the experimental samples was usually largest when the control was served first, but these differences were appreciably smaller than for the fresh control conditions. Note too, that serving the control last did not yield consistently higher or lower differences between control and experimental samples than did serving the control in a balanced order.

Discussion of Replications

Certain features about each of the replications raise doubts about the adequacy of the experiments. Each replication will be discussed in turn. Then we will draw what we feel are justified conclusions and suggest future research in this area.

Replication I

It was pointed out earlier that there is inadequate separation between the -80°C. and the -185°C. samples. In three experimental conditions, the former was rated higher than the latter, and in another three conditions the reverse was true, although none of the differences was statistically significant.

It was also shown that the best differentiation between the experimental samples was achieved with a fresh control served first or in balanced positions and with an irradiated control served last. (The latter fact would imply that the first three samples are best differentiated in terms of preference, the fourth contributing most to error). Reference to Table 3.4 shows that this conclusion holds true regardless of whether we consider the differences between the theoretically best and the theoretically poorest experimental samples, or whether we consider the maximum range among experimental samples in each experimental condition. The only qualification is that, in the latter case, the degree of differentiation among experimental samples seemed best for the irradiated control served last and next best for the fresh control served in a balanced order.

The fact that in the control-last situation the differentiation among samples depended heavily upon whether the control was fresh or irradiated is disturbing. In the control-last situation, unless one assumes that judges are clairvoyant, the differences should be equal or nearly so. Regardless of whether the control served last is fresh or irradiated, we are comparing the samples served in the first three positions only. Evidently something in this experiment went wrong. When the irradiated control was served last the maximum range of the experimental samples was 1.50, but was only .78 when the fresh control was last. Also, the difference between the theoretically best and the theoretically poorest samples was larger for the irradiated than for the fresh control in the last position (.89 vs .58). One might hypothesize various causes, such as non-equivalence of the samples from session to session or non-equivalence of judge-groups.

The analysis of variance also showed that the best differentiation between the control and experimental samples occurred when the fresh control was served either first or last. Fresh control served last is not the optimum condition for securing the best differentiation among experimental samples. One would not expect the differences between the variable control and these other samples, except for differences within the limits expected by sampling error and order effects. Taken at their face value (which should not be done), these data would indicate that by serving a fresh control in the very first position, one would achieve good differentiation among the experimental samples without sacrificing differentiation between the control and experimental samples. However, the data do not appear sufficiently "clean" to justify this conclusion.

Moreover, additional experimental conditions would have to be run to control for the position effects, that is, for the possibly higher rating of the fresh control due solely to the fact that it was served first.

Replication II

It was mentioned earlier that there was inadequate separation between the 4.5 Mrad and the 6.0 Mrad samples. The significance of the main effect of differences among experimental samples is largely attributable to the difference between the 3.0 Mrad and the other two samples.

Also, the differentiation between the experimental and control samples was better when the control was served in the first position (.93) than in the last position (.28); and the differentiation was poorest when the control was served in balanced order (.10). The significance of the three-factor interaction indicates that this conclusion depends in part upon the type of control. It holds primarily for the fresh control and less so, if at all, for the irradiated control.

The significance of the interaction between type of control and the difference between the experimental and control samples is trivial. The difference should be greater for the fresh control, since the irradiated control was physically the same as the experimental samples.

The most important factor is how well the experimental samples were differentiated from one another. As Table 3.2-II shows, something went wrong in the condition where the fresh control was served first. Here, the 3.0 Mrad sample and the 4.5 Mrad sample were rated lower than the 6.0 Mrad sample. The rating of the latter (6.22) seems too high. Moreover, the difference between the two extreme experimental conditions was 1.00 for the fresh control served last, but only .25 for the irradiated control served last. The two differences should be identical for reasons given in the discussion of Replication I.

There are also inconsistencies in position effects between the two replications. For example, in Replication I, the fresh control was rated highest when it was served last; in Replication II, it was rated highest when served first, and in both the balanced position yielded the lowest fresh control ratings.

Again, the data do not seem "clean" enough for any primary conclusions except that no evidence was obtained to indicate that a fresh control diminishes the differences among experimental samples.

Replication III

Because the experimental samples did not differ significantly from one another, also because of doubts about their uniformity, the findings from this replication are either trivial or inconclusive. Thus, it was shown that the control sample was significantly higher rated than were the experimental samples, that the difference was more pronounced when the control was fresh rather than irradiated (1.20 vs .07) and that the difference was greater when the control was served in the first position rather than in balanced order or in the last position (1.22, .49, and .20).

Another fact which raises doubt about this test is the high rating (7.19) for the irradiated control served first. This rating is nearly the same as for the fresh control served in the same position.

Replication IV

Although the average of the experimental samples differed from the control, again the 4.5 and 6.0 Mrad samples did not differ from one another. The analysis of variance shows that the difference between the experimental and control samples depended upon the type of control (.97 for fresh, .25 for irradiated), and the position of the control (1.19 for served first, .16 for served last, and .50 for balanced). As Table 3.4 shows, the greatest difference occurred when the fresh control was served first, and the next-largest when it was served in balanced order (1.87 and .89).

On the whole the results seem more consistent than the results in the other three replications.

3. Discussion & Conclusions

We have pointed out certain aspects of the data which raise doubts about the adequacy of the experiments. In the first two replications, the differences between the experimental samples appeared to be affected by the type of control even when served last; and the nature of the effect was inconsistent. In the first replication the combinations yielding the best differentiation were fresh control served either first or in balanced position and the irradiated control served last. In the second, the best combination was the fresh control served last. However, in no case were the two-factor interactions involving differentiation among experimental samples significant.

The most definite conclusion was that there was no good evidence that a fresh control per se affects the differentiation among experimental samples. Even this has to be tempered by the consideration that in one replication there were no significant differences among experimental samples, and that, in each of the others, two of the experimental samples did not differ significantly.

Any such lack of differences tends to support the null hypothesis. Obviously, if there were no real differences, their failure to emerge under certain conditions would be a trivial finding.

These experiments developed no evidence to indicate the advisability of changing the current standard testing practice wherein fresh controls are served in balanced order with experimental samples. The question of whether the position of the control affects the difference between the control and experimental samples was not answered. In three cases this interaction was significant, but it also represents a confounding of position of control with position effects in general. Additional control groups or experimental conditions would be needed to clarify the meaning of the finding.

Some of the data suggest that the position effects might be different if all samples were fresh than if all were irradiated. The results suggest, although they do not clearly indicate, that a fresh control is rated lowest when served in balanced order, but that an irradiated control is rated lowest when served last. One might hypothesize that poorer quality samples are affected more by the number of preceding samples since each one tends to increase the judge's awareness of any negative qualities present. The effect for good quality samples should be much less. If this hypothesis is true, then controlling for position effects becomes more complex since the effects (e.g., on awareness) differ for various qualities, and a universal "correction" is ruled out.

Another implication is somewhat more subtle. Typically, a fresh control is served with the irradiated samples. The error term for assessing the difference between the control and experimental samples is usually the judge-treatment interaction which is almost always lower than the between-judge variation; however, it can be used only when all judges rate both the control and the experimental samples. If the ratings of the fresh control remain fairly constant when it is served by itself the same number of times as are the experimental samples, while the ratings of the experimental samples decline in successive positions, then the difference between the control and the experimental samples would increase. It is possible that this increase could more than offset the higher error term (between-judges variation rather than judge-treatment interaction).

The entire area of order and position effects is relatively untapped in terms of both theory and research, and it is not too difficult to formulate and to provide the rationale for hypotheses.

The data in the present study are particularly heuristic, but the inconsistencies point to a fundamental problem whose resolution must precede further work. It is implied in the question "If we cannot replicate with basically similar types of foods, how can we generalize over many different types of foods?" One can only speculate timorously about the reasons for the inconsistencies. One is that conducting the study over a protracted interval of several months allowed something to happen to the test population. Perhaps the turnover in the subject pool brought about differences in the population preferences. Perhaps diurnal and seasonal factors induced systematic variability. Another possibility is that the food samples themselves were inconsistent. Their quality from lot to lot might have been so variable that differences in treatments were obscured or there may have been inversions of quality.

The implication of the first possibility is that we need to know more about the personal — psychological and physiological — factors related to preference, and more specifically, to the order of preference for various foods.

The next step would be to determine how combinations of these factors affect the validity of the tests, where validity is defined as the prediction of actual acceptance.

The major implication of the second possibility is that we might be trying too hard to measure an intrinsically unstable phenomenon. If food quality is that variable, then the importance of five adjustments of the levels of processing variables is open to question unless one can devise means of reducing the variability. This reduction is a matter of concern to the food technologist as well as to the behavioral scientist, with the work of the latter contingent upon the work of the former.

Although there is no definite proof as to what went wrong in these experiments, it would seem wise and practical in future methodological studies of this kind to use foods of greater physical uniformity and to conduct the entire study within a much shorter time span. Inter-session effects can be studied, and probably should be, since the foods themselves would be used under varying conditions. It would not be worthwhile to develop foods that had adequate acceptability only under certain restrictive conditions.

When this research project was initiated, we had hoped that analysis of the relationship between personal variables and food preferences (sampling topic) would shed light on correlates of preference. However, as was discussed in Chapter II, the relative homogeneity of the population and the lopsided distributions for many of the background characteristics would have attenuated any true underlying relationships.

Table 3.1. Characteristics of Each of the Four Replications

	Replication			
	I	II	III	IV
Product:	Roast Beef	Roast Beef	Ham	Ham
Processing variable	Temperature at time of irradiation	Irradiation dosage	Temperature at time of irradiation	Irradiation dosage
Levels of processing variable	Ambient -80°C. -185°C.	3.0 Mrad. 4.5 Mrad. 6.0 Mrad.	Ambient +5°C. -40°C.	3.5 Mrad 4.5 Mrad 6.0 Mrad
Constant processing factor	Irradiation dosage 3.5 Mrad	Temperature at time of irradiation -80°C.	Irradiation dosage 3.5 Mrad	Temperature at time of irradiation -40°C.
Was session confounded with experimental condition?	Yes	Yes	No*	No*

* All six conditions were present within each session

Table 3.2-I. Averages For Treatments According to Position and Type of Control - Replication I

			Experimental*				
Position	Type	Con-trol	Amb-ient	-80°C.	-185°C.	Ave. Exper.	Grand Ave.
First	Fresh	6.58	4.53	5.56	5.64	5.24	5.58
"	Irradiated	5.75	5.33	5.53	5.72	5.53	5.58
Last	Fresh	6.97	5.50	6.28	6.08	5.95	6.21
"	Irradiated	5.33	4.94	6.44	5.83	5.74	5.64
Balanced	Fresh	6.14	5.08	6.44	5.97	5.83	5.91
"	Irradiated	6.17	5.53	5.64	6.14	5.77	5.87
Position Totals							
	First	6.17	4.93	5.55	5.68	5.39	5.58
	Last	6.15	5.22	5.36	5.96	5.85	5.93
	Balanced	6.16	5.31	6.04	5.06	5.80	5.89
Type Totals							
	Fresh	6.56	5.04	6.09	5.90	5.67	5.90
	Irradiated	5.75	5.27	5.87	5.90	5.68	5.70
Grand Total		6.16	5.15	5.98	5.90	5.68	5.80

* All irradiated at 3.5 Mrads.

Table 3.2-II. Averages For Treatments According to Position and Type of Control - Replication II

			Experimental*				
Position	Type	Con-trol	3.0 Mrad	4.5 Mrad	6.0 Mrad	Ave. Exper.	Grand Ave.
First	Fresh	7.39	5.81	5.50	6.22	5.84	6.23
"	Irradiated	6.14	6.22	5.79	5.50	5.84	5.91
Last	Fresh	6.89	6.67	6.11	5.67	6.15	6.34
"	Irradiated	5.75	6.31	5.44	6.06	5.94	5.89
Balanced	Fresh	6.44	6.39	6.36	6.11	6.29	6.33
"	Irradiated	6.08	6.28	5.72	6.06	6.02	6.04
Position Totals							
	First	6.77	6.02	5.65	5.86	5.84	6.08
	Last	6.32	6.49	5.78	5.87	6.05	6.12
	Balanced	6.26	6.34	6.04	6.09	6.16	6.18
Type Totals							
	Fresh	6.91	6.29	5.99	6.00	6.09	6.30
	Irradiated	5.99	6.27	5.65	5.87	5.94	5.95
Grand Total		6.45	6.28	5.82	5.94	6.01	6.10

* All irradiated at -80°C.

Table 3.2-III. Averages For Treatments According to Position and Type of Control - Replication III

			Experimental*				
Position	Type	Con-trol	Amb-ient	+5°C.	-40°C.	Ave. Exper.	Grand Ave.
First	Fresh	7.39	5.81	5.42	5.92	5.72	6.14
"	Irradiated	7.19	6.22	6.56	6.44	6.41	6.60
Last	Fresh	7.17	6.28	6.53	6.56	6.46	6.64
"	Irradiated	6.08	6.25	6.39	6.53	6.39	6.31
Balanced	Fresh	7.08	5.94	5.72	5.89	5.85	6.16
"	Irradiated	6.00	5.75	6.50	6.47	6.24	6.18
Position Totals							
	First	7.29	6.02	5.99	6.18	6.07	6.37
	Last	6.63	6.27	6.46	6.55	6.43	6.48
	Balanced	6.54	5.85	6.11	6.18	6.05	6.17
Type Totals							
	Fresh	7.21	6.01	5.89	6.12	6.01	6.31
	Irradiated	6.42	6.07	6.48	6.48	6.35	6.36
Grand Total		6.82	6.05	6.19	6.30	6.18	6.34

* All irradiated at 3.5 Mrads.

Table 3.2-IV. Averages For Treatments According to Position and Type of Control - Replication IV

			Experimental*				
Position	Type	Con-trol	3.5 Mrad	4.5 Mrad	6.0 Mrad	Ave. Expr.	Grand Ave.
First	Fresh	7.31	5.81	5.11	5.39	5.44	5.91
"	Irradiated	6.33	6.14	5.78	5.50	5.81	5.94
Last	Fresh	6.36	6.36	6.17	6.06	6.19	6.24
"	Irradiated	5.81	6.08	5.47	5.44	5.66	5.70
Balanced	Fresh	6.44	5.72	5.39	5.53	5.55	5.77
"	Irradiated	6.17	6.42	5.72	6.14	6.09	6.11
Position Totals							
	First	6.82	5.98	5.45	5.45	5.63	5.93
	Last	6.09	6.21	5.82	5.75	5.93	5.97
	Balanced	6.32	6.07	5.56	5.84	5.82	5.95
Type Totals							
	Fresh	6.70	5.96	5.56	5.66	5.73	5.97
	Irradiated	6.10	6.21	5.66	5.69	5.85	5.92
Grand Total		6.40	6.09	5.61	5.68	6.79	5.95

*All irradiated at -40° C.

Table 3.3. Analyses of Variance for the Four Replications

Source of Variation		Replication I		Replication II	
	df	F	Sig	F	Sig.
(A) Position of Control (First,last,balanced)	2	1.26	—	<1	—
(B) Type of Control (Fresh,Irradiated)	1	1.07	—	3.93	5%
(A) X (B)	2	<1	—	<1	—
(C_1) Among Experimental Samples	2	21.18	.1%	8.13	.1%
(C_2) Between Experimental and Control Samples	1	17.77	.1%	20.61	.1%
(D) Among Judges (Within (A) and (B)	210	ms=	8.20	ms=	6.80
(A) X (C_1)	4	<1	—	<1	—
(A) X (C_2)	2	1.78	—	6.63	1%
(B) X (C_1)	2	1.32	—	<1	—
(B) X (C_2)	1	12.88	.1%	15.21	.1%
(A) X (B) X (C_1)	4	2.62	5%	4.20	1%
(A) X (B) X (C_2)	2	4.16	1%	3.15	5%
(C) X (D) (Within (A) and (B))	630	ms=	2.10	ms=	1.51

Table 3.3. Analyses of Variance for the Four Replications (Continued)

Source of Variation		Replication III		Replication IV	
	df	F	Sig.	F	Sig.
(A) Position of Control (First, last, balanced)	2	< 1	—	< 1	—
(B) Type of Control (Fresh, Irradiated)	1	< 1	—	< 1	—
(A) X (B)	2	< 1	—	1.48	—
(C_1) Among Experimental Samples	2	2.13	—	8.61	.1%
(C_2) Between Experimental and Control Samples	1	39.23	.1%	35.76	.1%
(D) Among Judges (Within (A) and (B)	210	ms=	8.36	ms=	9.60
(A) X (C_1)	4	< 1	—	< 1	—
(A) X (C_2)	2	8.86	.1%	9.08	.1%
(B) X (C_1)	2	2.22	—	< 1	—
(B) X (C_2)	1	30.06	.1%	12.75	.1%
(A) X (B) X (C_1)	4	1.10	—	< 1	—
(A) X (B) X (C_2)	2	< 1	—	3.48	5%
(C) X (D) (Within (A) and (B))	630	ms=	1.71	ms=	1.70

NOTE: (A), (B), and (A) X (B) were each tested against (D). All other effects were tested against (C) X (D).

Table 3.4. Algebraic Differences In Mean Rating Between Samples Related to Position and Type of Control Sample

	A. Best minus poorest experimental*		B. Maximum range of experimental		C. Control minus average experimental	
	Fresh Control	Irrad. Control	Fresh Control	Irrad. Control	Fresh Control	Irrad. Control
Replication I						
Position of Control:						
First	1.11	.39	1.11	.39	1.34	.22
Last	.58	.89	.78	1.50	1.02	-.41
Balanced	.89	.61	1.36	.61	.31	.40
Replication II						
Position of Control:						
First	-.41	.72	.72	.72	1.55	.30
Last	1.00	.25	1.00	.87	.74	-.19
Balanced	.28	.22	.28	.56	.15	.06
Replication III						
Position of Control:						
First	.11	.22	.50	.34	1.67	.78
Last	.28	.28	.28	.28	.71	-.31
Balanced	-.05	.72	.22	.75	1.23	-.24
Replication IV						
Position of Control						
First	.42	.64	.70	.64	1.87	.52
Last	.30	.64	.30	.64	.17	.15
Balanced	.19	.28	.33	.70	.89	.08

*Theoretically best minus theoretically poorest treatments

CHAPTER V. GENERAL SUMMARY AND RECOMMENDATIONS

The three topics of this contract represented an ambitious undertaking designed to increase the validity, reliability, and efficiency of sensory evaluation methods. The original scope, while not coextensive with the range of sensory evaluation problems, was extremely comprehensive. The execution of the original plans depended heavily upon the cooperation of many elements of the U.S. Army Natick Laboratories. The absence of cooperation anywhere along the chain usually meant the inability to carry out the plans. Such absence did not imply lack of willingness or of technical proficiency. Both of these were present, however, commitments to other projects and personnel turnover were obstacles in the quest for a viable methodological research program.

The experience from this contract suggests that much more in-house directed and implemented research is advisable. It is difficult to outline on a priori grounds each of a series of experiments. A sounder approach seems to be the sequential one, building on the past and modifying the experimental plans in light of preceding work. Certainly, this suggestion does not preclude the use of outside consultants; but it does imply that an internally fused and fueled program, which takes advantage of the experimenters' observations of the details and intricacies of the broader food research and development effort, is the type of program which is most likely to be productive.

Some light has been shed on several concrete problems — number of samples served and contrast and convergence. Recommendations for further research were made, and there were observations about the possible lack of quality uniformity among experimental food samples. The data presented here can serve as the nucleus for further and more intensive research, since we now know better where future research effort should be concentrated and the nature of the hurdles to overcome.

Methodological research cannot be a stepchild. Of course, there is a pressing need for "production-type" sensory evaluation tests; these tests are ultimately the raison d'etre. At the same time there is the obligation to insure that such tests are as sound — leading to valid conclusions and recommendations — as is fitting their purpose. Improvements in testing methodology should be continous and will be if satisfaction with the status quo can be avoided.

REFERENCES

Bradley, J.E. "Influence of continued testing on preference ratings". In David R. Peryam, Franci. J. Pilgrim, and Martin S. Peterson (Edictors), Food Acceptance Testing Methodology, A Symposium. National Academy of Sciences — National Research Council, October, 1954. (Quartermaster Food and Container Institute for the Armed Forces, Chicago, Illinois).

Eindhoven, J., Peryam, D.R., Heiligman, F., and Hamman, J.W. Effects of sample sequence on food preferences. Journal of Food Science, 1965, 29, 520-524.

Kamenetzky, J.M. Contrast and convergence effects in the ratings of foods. Journal of Applied Psychology, 1959, 43, 47-52.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)*

Peryam & Kroll Research Corporation
Chicago, Illinois

2a REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b GROUP

3 REPORT TITLE

"Studies on Acceptance Testing Methodology: Preliminary Studies on Characteristics of Taste Panel, Sample Size, and Contrast and Convergence Effects"

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5 AUTHOR(S) *(Last name, first name, initial)*

Kamen Joseph M. Peryam, David R.
Peryam, David B. Kroll, Beverley J.

6 REPORT DATE July 1967	7a TOTAL NO OF PAGES 65	7b NO OF REFS 3

8a CONTRACT OR GRANT NO

DA-19-129-AMC-734(N)

b PROJECT NO

1T024701A121-02

c

d

9a ORIGINATOR'S REPORT NUMBER(S)

68-10-PR

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10 AVAILABILITY/LIMITATION NOTICES

This document has been approved for public release and sale; its distribution is unlimited.

11 SUPPLEMENTARY NOTES

12 SPONSORING MILITARY ACTIVITY

United States Army Natick Laboratories
Natick, Massachusetts 01760

13 ABSTRACT

This report summarizes research accomplished in methodology aspects of se sory evaluation testing. It also discusses certain studies which were designed, but, for administrative reasons, could not be completed.

Two main studies are presented in detail. The first investigated the effect of the number of samples upon differences in preference between selected samples as a function of whether they were included in the first half or second half of a series. There was no evidence that the number of samples -- varying from 2 to 12 -- had any consistent or significant effect on preference differences; however, the data do suggest several hypothesis for future investigation.

The second study attempted primarily to determine the effect of a fresh vs an irradiated control on preference differences, among various irradiated samples. There were logical inconsistencies in the data; however, there was no basis for concluding that a fresh control attenuates the differences in preference. It appeared that quality control of the test products needed to be tightered. Methodological research in this area are discussed. Recommendations for further in-house work by the U. S. Army Natick Laboratories are made.

The report also covers the topic of sampling test subjects in-house, describes the panel population in ~~the~~ U. S. Army Natick Laboratories, and points our certain interrelationships among the panel member's background characteristics. These date suggest certain points for future investigation.

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Measurement	8		9			
Evaluation	8		8			
Food Preferences	8,9		9			
Taste Tests	10		10			
Military Requirements	4		4			
Military Rations	4		4			
Development			8			
Methods			9			

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
_______________________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
_______________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
_______________________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.